INTERNATIONAL STANDARD

ISO/IEC 13818-2

Second edition
2000-12-15

Information technology — Generic coding of moving pictures and associated audio information: Video

Technologies de l'information — Codage générique des images animées et du son associé: Données vidéo

Reference number
ISO/IEC 13818-2:2000(E)

Intro. 4 The scalable and the non-scalable syntax

The full syntax can be divided into two major categories: One is the non-scalable syntax, which is structured as a super set of the syntax defined in ISO/IEC 11172-2. The main feature of the non-scalable syntax is the extra compression tools for interlaced video signals. The second is the scalable syntax, the key property of which is to enable the reconstruction of useful video from pieces of a total bitstream. This is achieved by structuring the total bitstream in two or more layers, starting from a standalone base layer and adding a number of enhancement layers. The base layer can use the non-scalable syntax, or in some situations conform to the ISO/IEC 11172-2 syntax.

Intro. 4.1 Overview of the non-scalable syntax

The coded representation defined in the non-scalable syntax achieves a high compression ratio while preserving good image quality. The algorithm is not lossless as the exact sample values are not preserved during coding. Obtaining good image quality at the bitrates of interest demands very high compression, which is not achievable with intra picture coding alone. The need for random access, however, is best satisfied with pure intra picture coding. The choice of the techniques is based on the need to balance a high image quality and compression ratio with the requirement to make random access to the coded bitstream.

A number of techniques are used to achieve high compression. The algorithm first uses block-based motion compensation to reduce the temporal redundancy. Motion compensation is used both for causal prediction of the current picture from a previous picture, and for non-causal, interpolative prediction from past and future pictures. Motion vectors are defined for each 16-sample by 16-line region of the picture. The prediction error, is further compressed using the Discrete Cosine Transform (DCT) to remove spatial correlation before it is quantised in an irreversible process that discards the less important information. Finally, the motion vectors are combined with the quantised DCT information, and encoded using variable length codes.

Intro. 4.1.1 Temporal processing

Because of the conflicting requirements of random access and highly efficient compression, three main picture types are defined. Intra Coded Pictures (I-Pictures) are coded without reference to other pictures. They provide access points to the coded sequence where decoding can begin, but are coded with only moderate compression. Predictive Coded Pictures (P-Pictures) are coded more efficiently using motion compensated prediction from a past intra or predictive coded picture and are generally used as a reference for further prediction. Bidirectionally-predictive Coded Pictures (B-Pictures) provide the highest degree of compression but require both past and future reference pictures for motion compensation. Bidirectionally-predictive coded pictures are never used as references for prediction (except in the case that the resulting picture is used as a reference in a spatially scalable enhancement layer). The organisation of the three picture types in a sequence is very flexible. The choice is left to the encoder and will depend on the requirements of the application. Figure Intro. 1 illustrates an example of the relationship among the three different picture types.

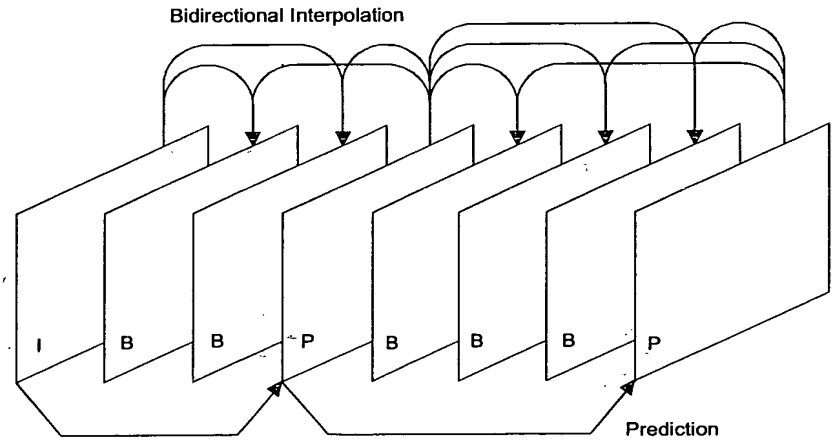


Figure Intr . 1 – Example of temporal picture structure

Intro. 4.1.2 Coding interlaced video

Each frame of interlaced video consists of two fields which are separated by one field-period. The Specification allows either the frame to be encoded as picture or the two fields to be encoded as two pictures. Frame encoding or field encoding can be adaptively selected on a frame-by-frame basis. Frame encoding is typically preferred when the video scene contains significant detail with limited motion. Field encoding, in which the second field can be predicted from the first, works better when there is fast movement.

Intro. 4.1.3 Motion representation – Macroblocks

As in ISO/IEC 11172-2, the choice of 16 by 16 macroblocks for the motion-compensation unit is a result of the trade-off between the coding gain provided by using motion information and the overhead needed to represent it. Each macroblock can be temporally predicted in one of a number of different ways. For example, in frame encoding, the prediction from the previous reference frame can itself be either frame-based or field-based. Depending on the type of the macroblock, motion vector information and other side information is encoded with the compressed prediction error in each macroblock. The motion vectors are encoded differentially with respect to the last encoded motion vectors using variable length codes. The maximum length of the motion vectors that may be represented can be programmed, on a picture-by-picture basis, so that the most demanding applications can be met without compromising the performance of the system in more normal situations.

It is the responsibility of the encoder to calculate appropriate motion vectors. This Specification does not specify how this should be done.

Intro. 4.1.4 Spatial redundancy reduction

Both source pictures and prediction errors have high spatial redundancy. This Specification uses a block-based DCT method with visually weighted quantisation and run-length coding. After motion compensated prediction or interpolation, the resulting prediction error is split into 8 by 8 blocks. These are transformed into the DCT domain where they are weighted before being quantised. After quantisation many of the DCT coefficients are zero in value and so two-dimensional run-length and variable length coding is used to encode the remaining DCT coefficients efficiently.

Intro. 4.1.5 Chrominance formats

In addition to the 4:2:0 format supported in ISO/IEC 11172-2 this Specification supports 4:2:2 and 4:4:4 chrominance formats.

Intro. 4.2 Scalable extensions

The scalability tools in this Specification are designed to support applications beyond that supported by single layer video. Among the noteworthy applications areas addressed are video telecommunications, video on Asynchronous Transfer Mode (ATM) networks, interworking of video standards, video service hierarchies with multiple spatial, temporal and quality resolutions, HDTV with embedded TV, systems allowing migration to higher temporal resolution HDTV, etc. Although a simple solution to scalable video is the simulcast technique which is based on transmission/storage of multiple independently coded reproductions of video, a more efficient alternative is scalable video coding, in which the bandwidth allocated to a given reproduction of video can be partially re-utilised in coding of the next reproduction of video. In scalable video coding, it is assumed that given a coded bitstream, decoders of various complexities can decode and display appropriate reproductions of coded video. A scalable video encoder is likely to have increased complexity when compared to a single layer encoder. However, this Recommendation | International Standard provides several different forms of scalabilities that address non-overlapping applications with corresponding complexities. The basic scalability tools offered are:

- data partitioning;
- SNR scalability;
- spatial scalability; and
- temporal scalability.

Moreover, combinations of these basic scalability tools are also supported and are referred to as *hybrid scalability*. In the case of basic scalability, two layers of video referred to as the *lower layer* and the *enhancement layer* are allowed, whereas in hybrid scalability up to three layers are supported. Tables Intro. 1 to Intro. 3 provide a few example applications of various scalabilities.

6.1.1.2 Frame

A frame consists of three rectangular matrices of integers: a luminance matrix (Y), and two chrominance matrices (Cb and Cr).

The relationship between these Y, Cb and Cr components and the primary (analogue) Red, Green and Blue Signals (E'_R, E'_G and E'_B), the chromaticity of these primaries and the transfer characteristics of the source frame may be specified in the bitstream (or specified by some other means). This information does not affect the decoding process.

6.1.1.3 Field

A field consists of every other line of samples in the three rectangular matrices of integers representing a frame.

A frame is the union of a top field and a bottom field. The top field is the field that contains the top-most line of each of the three matrices. The bottom field is the other one.

6.1.1.4 Picture

A coded picture is made of a picture header, the optional extensions immediately following it and the following picture data. A coded picture may be a coded frame or a coded field.

An I-frame picture or a pair of field pictures, where the first field picture is an I-picture and the second field picture is an I-picture or a P-picture, is called a coded I-frame.

A P-frame picture or a pair of P-field pictures is called a coded P-frame.

A B-frame picture or a pair of B-field pictures is called a coded B-frame.

A coded I-frame, a coded P-frame or a coded B-frame is called a coded frame.

A reconstructed picture is obtained by decoding a coded picture, i.e. a picture header, the optional extensions immediately following it, and the picture data. A coded picture may be a frame picture or a field picture. A reconstructed picture is either a reconstructed frame (when decoding a frame picture), or one field of a reconstructed frame (when decoding a field picture).

6.1.1.4.1 Field pictures

If field pictures are used, then they shall occur in pairs (one top field followed by one bottom field, or one bottom field followed by one top field) and together constitute a coded frame. The two field pictures that comprise a coded frame shall be encoded in the bitstream in the order in which they shall occur at the output of the decoding process.

When the first picture of the coded frame is a P-field picture, then the second picture of the coded frame shall also be a P-field picture. Similarly when the first picture of the coded frame is a B-field picture the second picture of the coded frame shall also be a B-field picture.

When the first picture of the coded frame is a I-field picture, then the second picture of the frame shall be either an I-field picture or a P-field picture. If the second picture is a P-field picture, then certain restrictions apply (see 7.6.3.5).

6.1.1.4.2 Frame pictures

When coding interlaced sequences using frame pictures, the two fields of the frame shall be interleaved with one another and then the entire frame is coded as a single frame-picture.

6.1.1.5 Picture types

There are three types of pictures that use different coding methods:

- an **Intra-coded (I) picture** is coded using information only from itself;
- a **Predictive-coded (P) picture** is a picture which is coded using motion compensated prediction from a past reference frame or past reference field;
- a **Bidirectionally predictive-coded (B) picture** is a picture which is coded using motion compensated prediction from a past and/or future reference frame(s).

6.1.1.6 Sequence header

A video sequence header commences with a sequence_header_code and is followed by a series of data elements. In this Specification, sequence_header() shall be followed by sequence_extension() which includes further parameters beyond those used by ISO/IEC 11172-2. When sequence_extension() is present, the syntax and semantics defined in ISO/IEC 11172-2 do not apply, and the present Specification applies.

6.1.1.9 4:2:2 format

In this format the Cb and Cr matrices shall be one half the size of the Y-matrix in the horizontal dimension and the same size as the Y-matrix in the vertical dimension. The Y-matrix shall have an even number of samples.

NOTE – When interlaced frames are coded as field pictures, the picture reconstructed from each of these field pictures shall have a Y-matrix with half the number of lines as the corresponding frame. Thus the total number of lines in the Y-matrix of an entire frame shall be divisible by two.

The luminance and chrominance samples are positioned as shown in Figure 6-5.

In order to clarify the organisation, Figure 6-6 shows the (vertical) positioning of the samples when the frame is separated into two fields.

6.1.1.10 4:4:4 format

In this format the Cb and Cr matrices shall be the same size as the Y-matrix in the horizontal and the vertical dimensions.

NOTE – When interlaced frames are coded as field pictures, the picture reconstructed from each of these field pictures shall have a Y-matrix with half the number of lines as the corresponding frame. Thus the total number of lines in the Y-matrix of an entire frame shall be divisible by two.

The luminance and chrominance samples are positioned as shown in Figures 6-6 and 6-7.

6.1.1.11 Frame re-ordering

When the sequence contains coded B-frames, the number of consecutive coded B-frames is variable and unbounded. The first coded frame after a sequence header shall not be a B-frame.

A sequence may contain no coded P-frames. A sequence may also contain no coded I-frames in which case some care is required at the start of the sequence and within the sequence to effect both random access and error recovery.

The order of the coded frames in the bitstream, also called coded order, is the order in which a decoder reconstructs them. The order of the reconstructed frames at the output of the decoding process, also called the display order, is not always the same as the coded order and this subclause defines the rules of frame re-ordering that shall happen within the decoding process.

When the sequence contains no coded B-frames, the coded order is the same as the display order. This is true in particular always when low_delay is one.

When B-frames are present in the sequence, re-ordering is performed according to the following rules:

- If the current frame in coded order is a B-frame, the output frame is the frame reconstructed from that B-frame.
- If the current frame in coded order is a I-frame or P-frame, the output frame is the frame reconstructed from the previous I-frame or P-frame if one exists. If none exists, at the start of the sequence, no frame is output.

The frame reconstructed from the final I-frame or P-frame is output immediately after the frame reconstructed when the last coded frame in the sequence was removed from the VBV buffer.

The following is an example of frames taken from the beginning of a video sequence. In this example there are two coded B-frames between successive coded P-frames and also two coded B-frames between successive coded I- and P-frames and all pictures are frame-pictures. Frame '1I' is used to form a prediction for frame '4P'. Frames '4P' and '1I' are both used to form predictions for frames '2B' and '3B'. Therefore the order of coded frames in the coded sequence shall be '1I', '4P', '2B', '3B'. However, the decoder shall display them in the order '1I', '2B', '3B', '4P'.

At the encoder input:

1	2	3	4	5	6	7	8	9	10	11	12	13
I	B	B	P	B	B	P	B	B	I	B	B	P

At the encoder output, in the coded bitstream, and at the decoder input:

1	4	2	3	7	5	6	10	8	9	13	11	12
I	P	B	B	P	B	B	I	B	B	P	B	B

At the decoder output:

1	2	3	4	5	6	7	8	9	10	11	12	13

6.2.2.6 Group of pictures header

group_of_pictures_header() {	No. of bits	Mnemonic
group_start_code	32	bslbf
time_code	25	uimsbf
closed_gop	1	uimsbf
broken_link	1	uimsbf
next_start_code()		
}		

6.2.3 Picture header

picture_header() {	No. of bits	Mnemonic
picture_start_code	32	bslbf
temporal_reference	10	uimsbf
picture_coding_type	3	uimsbf
vbv_delay	16	uimsbf
if (picture_coding_type == 2 || picture_coding_type == 3) {		
full_pel_forward_vector	1	bslbf
forward_f_code	3	bslbf
}		
if (picture_coding_type == 3) {		
full_pel_backward_vector	1	bslbf
backward_f_code	3	bslbf
}		
while (nextbits() == '1') {		
extra_bit_picture /* with the value '1' */	1	uimsbf
extra_information_picture	8	uimsbf
}		
extra_bit_picture /* with the value '0' */	1	uimsbf
next_start_code()		
}		

This subclause does not adequately document the block layer syntax when data partitioning is used. See 7.10.

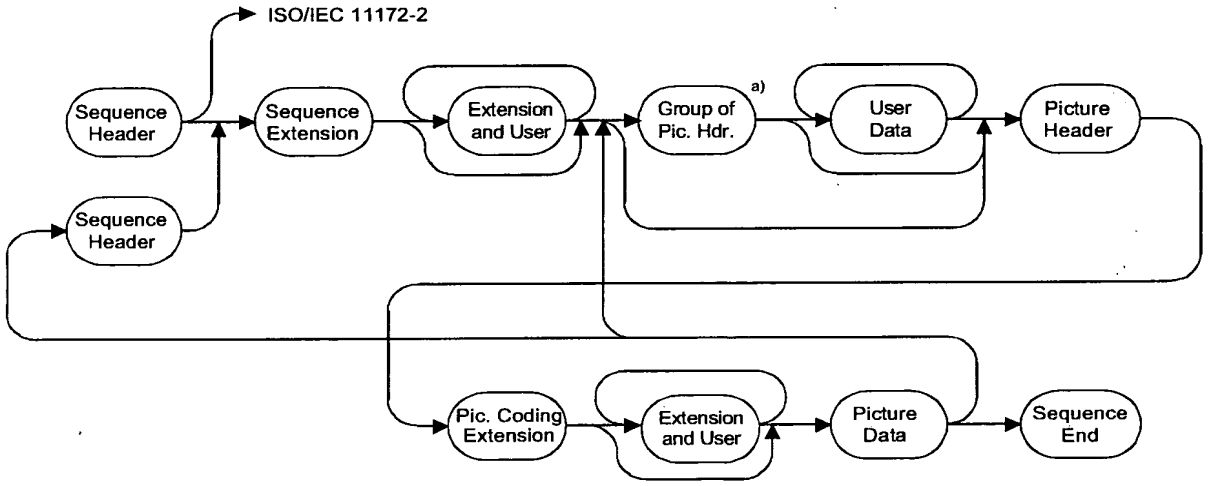
block(i) {	No. of bits	Mnemonic
if (pattern_code[i]) {		
if (macroblock_intra) {		
if (i < 4) {		
dct_dc_size_luminance	2-9	vlclbf
if(dct_dc_size_luminance != 0)		
dct_dc_differential	1-11	uimsbf
} else {		
dct_dc_size_chrominance	2-10	vlclbf
if(dct_dc_size_chrominance != 0)		
dct_dc_differential	1-11	uimsbf
}		
} else {		
First DCT coefficient	2-24	vlclbf
}		
while (nextbits() != End of block)		
Subsequent DCT coefficients	3-24	vlclbf
End of block	2 or 4	vlclbf
}		
}		

6.3 Video bitstream semantics

6.3.1 Semantic rules for higher syntactic structures

This subclause details the rules that govern the way in which the higher level syntactic elements may be combined together to produce a legal bitstream. Subsequent clauses detail the semantic meaning of all fields in the video bitstream.

Figure 6-15 illustrates the high level structure of the video bitstream.



a) After a GOP, the first picture shall be an I-picture.

Figure 6-15 – High level bitstream organisation

6.3.9 Picture header

picture_start_code – The picture_start_code is a string of 32 bits having the value 00000100 in hexadecimal.

temporal_reference – The temporal_reference is a 10-bit unsigned integer associated with each coded picture.

The following simple specification applies only when low_delay is equal to zero.

When a coded frame is in the form of two field pictures, the temporal_reference associated with each picture shall be the same (it is called the temporal_reference of the coded frame). The temporal_reference of each coded frame shall increment by one modulo 1024 when examined in display order at the output of the decoding process, except when a group of pictures header occurs. Among the frames coded after a group of pictures header, the temporal_reference of the coded frame that is displayed first, shall be set to zero.

The following more general specification applies when low_delay is equal to zero or one.

If picture A is not a big picture, i.e. the VBV buffer is only examined once before the coded picture A is removed from the VBV buffer and if N is the temporal_reference of picture A, then the temporal_reference of picture B immediately following picture A in display order is equal to:

- 0 if there is a group of pictures header present between picture A and picture B (in coded order).
- (N + 1) % 1024 if picture B is a frame picture or is the first field of a pair of field pictures.
- N if picture B is the second field of a pair of field pictures.

When low_delay is equal to one, there may be situations where the VBV buffer shall be re-examined several times before removing a coded picture (referred to as a big picture) from the VBV buffer.

If picture A is a big picture and if K is the number of times that the VBV buffer is re-examined as defined in C.7 (K > 0), if N is the temporal_reference of picture A, then the temporal_reference of picture B immediately following picture A in display order is equal to:

- K % 1024 if there is a group of pictures header present between picture A and picture B (in coded order).
- (N + K + 1) % 1024 if picture B is a frame picture or is the first field of a pair of field pictures.
- (N + K) % 1024 if picture B is the second field of a pair of field pictures.

NOTE 1 – If the big picture is the first field of a frame coded with field pictures, then the temporal_reference of the two field pictures of that coded frame are not identical.

picture_coding_type – The picture_coding_type identifies whether a picture is an intra-coded picture(I), predictive-coded picture(P) or bidirectionally predictive-coded picture(B). The meaning of picture_coding_type is defined in Table 6-12.

NOTE 2 – Intra-coded pictures with only DC coefficients (D-pictures) that may be used in ISO/IEC 11172-2 are not supported by this Specification.

Table 6-12 – picture_coding_type

picture_coding_type	coding method
000	Forbidden
001	intra-coded (I)
010	predictive-coded (P)
011	bidirectionally-predictive-coded (B)
100	Shall not be used (dc intra-coded (D) in ISO/IEC11172-2)
101	Reserved
110	Reserved
111	Reserved

7 The video decoding process

This clause specifies the decoding process that a decoder shall perform to reconstruct frames from the coded bitstream.

The IDCT function f[y][x] used in the decoding process may be any of several approximations of the saturated mathematical integer-number IDCT defined in Annex A. Requirements on the accuracy of the IDCT function used in the decoding process are specified in Annex A.

In 7.1 through 7.6 the simplest decoding process is specified in which no scalability features are used. Subclauses 7.7 to 7.11 specify the decoding process when scalable extensions are used. Subclause 7.12 defines the output of the decoding process.

Figure 7-1 is a diagram of the Video Decoding Process without any scalability. The diagram is simplified for clarity.

NOTE – Throughout this Specification two dimensional arrays are represented as *name*[*q*][*p*] where '*q*' is the index in the vertical dimension and '*p*' the index in the horizontal dimension.

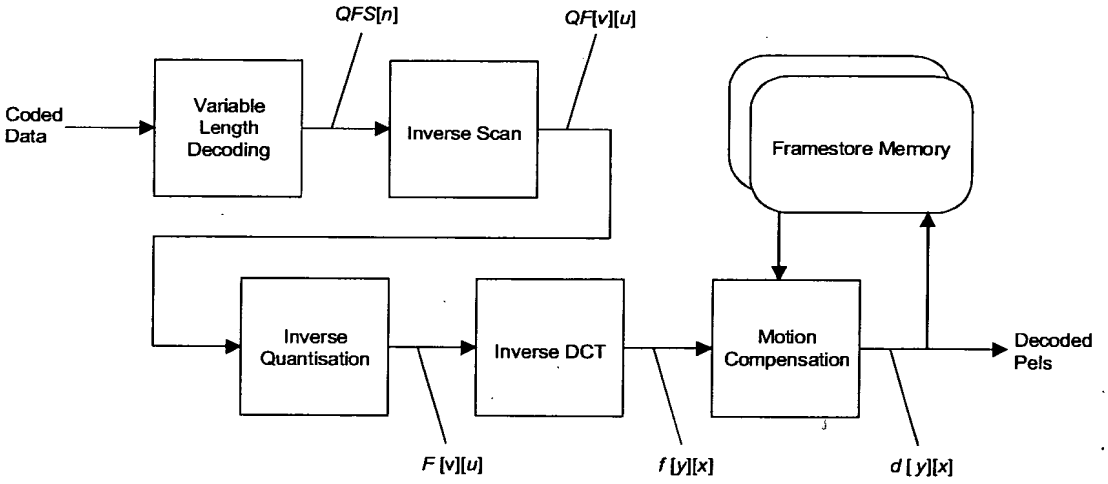


Figure 7-1 – Simplified Video Decoding Process

7.1 Higher syntactic structures

The various parameters and flags in the bitstream for macroblock() and all syntactic structures above macroblock() shall be interpreted as indicated in clause 6. Many of these parameters and flags affect the decoding process described in the following subclauses. Once all of the macroblocks in a given picture have been processed, the entire picture will have been reconstructed.

Reconstructed fields shall be associated together in pairs to form reconstructed frames. (See "picture_structure" in 6.3.10.)

The sequence of reconstructed frames shall be re-ordered as described in 6.1.1.11.

If progressive_sequence == 1 the reconstructed frames shall be output from the decoding process at regular intervals of the frame period as shown in Figure 7-19.

If progressive_sequence == 0 the reconstructed frames shall be broken into a sequence of fields which shall be output from the decoding process at regular intervals of the field period as shown in Figure 7-20. In the case that a frame picture has repeat_first_field == 1 the first field of the frame shall be repeated after the second field. (See "repeat_first_field" in 6.3.10.)